# ON OPTIMIZING COMPUTATIONS FOR TRANSITION MATRICES

R. E. McFarland and A. B. Rochkind

Reprinted from IEEE Transactions on Automatic Control, Vol. AC-23, No. 3, June 1978
0018-9286/78/0600-0495$00.75 

# On Optimizing Computations for Transition Matrices

R. E. McFARLAND AND A. B. ROCHKIND

*Abstract*—**For the special case where the coefficient matrix is in standard companion form, all of the elements of the transition matrix may be obtained recursively from a single row of elements. The number of computational steps necessary to generate this required row is an order of magnitude less than that required for general coefficient matrices. Also the forced-response coefficients are shown to require negligible additional calculations for the common problem with a single input and a zero-order data hold. These computational savings enable the typical, modest-sized digital computer to address the previously formidable problem of nonstationary, high-order transfer functions in real time.**

## INTRODUCTION

Consider the single-input stationary system of order $n$ defined by

$$\dot{X}(t) = AX(t) + fu(t) \tag{1}$$

where $X(t) = [x_1(t), \cdots, x_n(t)]'$, $A$ is an $n^2$ matrix, $f$ is an $n$ vector, and $u(t)$ is a scalar input function. This system, which is often encountered in control theory applications, may be solved by algorithms which have recently been devised [1] requiring $0(n^3)$ multiplicative operations. However, considerable simplification and reduction in computational effort results if the system is known to be in phase canonical form

$$\dot{X}(t) = AX(t) + U(t) \tag{2}$$

where

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ -a_1 & -a_2 & -a_3 & \cdots & -a_n \end{bmatrix} \tag{3}$$

and

$$U(t) = [0, 0, \cdots, u(t)]'. \tag{4}$$

Systems given by (1) not in canonical form can be transformed into that form if they are controllable [2]. This transformation, of course, requires the solution to the characteristic equation of $A$ [3], which

Manuscript received June 24, 1977. This work was performed while both authors were members of Computer Sciences Corporation, under contract to the Simulation Sciences Division of NASA, Ames Research Center.

R. E. McFarland is with the U.S. Army Aviation R&D Command at Ames Research Center, Moffett Field, CA, under the direction of the Simulation Sciences Division.

A. B. Rochkind is with Spectra Physics, Mountain View, CA.

0018-9286/78/0600-0495$00.75 

therefore requires $0(n^3)$ multiplications [4]. However, the system will be naturally in canonical form if it arises from either the conversion of an equation of order $n$ to a system of first-order equations or if it represents a transfer function in state-space form.

The transfer-function application is of special importance for real-time simulation in the case of nonstationary systems. In this problem, it is assumed that the digital technology exists to implement transition intervals which are small enough so that coefficient derivatives may be ignored in the treatment of the discrete solution to (2) as a recursive boundary value problem [5], [6].

The discrete solution of (2) requires the computation of the state transition matrix. Applications of the Cayley–Hamilton theorem [7]–[10] have indicated the path to considerable reductions in the number of operations required to produce the transient response ($u(t)\equiv 0$). This note extends these results to the computation of the forced response with negligible additional operations, providing some data-hold assumption is made on the input function during the transition interval. For brevity only the zero-order hold is discussed here. It will be shown that for systems in canonical form the computation of the entire forced response requires only $0(n^2)$ multiplications.

A coefficient sizing parameter is developed for the estimation of the number of necessary terms for series convergence.

## Discussion

For a single interval of transition $h$, assuming a zero-order hold, the discrete solution to (2) may be written as [11].

$$X(h) = T_0 X(0) + T_1 U(0) \tag{5}$$

where

$$T_m = h^m \sum_{j\geq 0} \frac{h^j A^j}{(j+m)!}. \tag{6}$$

The $T_m$ are called *generalized transition matrices*, and $T_0$ is the state transition matrix. Note by (4) that only the last column of $T_1$ is required in (5). For any $m$ (6) may be expressed as a recurrence relationship

$$T_{m-1} = AT_m + I\frac{h^{m-1}}{(m-1)!} \tag{7}$$

where $I$ is the identity matrix. From (3) and (7)

$$T_{m-1}(i,j) = T_m(i+1,j) + \frac{h^{m-1}\delta_{i,j}}{(m-1)!} \quad \begin{cases} (1\leq i<n) \\ (1\leq j\leq n) \end{cases} \tag{8}$$

where $\delta_{i,j}$ is the Kronecker delta. Specializing (8) to $j=n$

$$T_m(i,n) = T_{m-1}(i-1,n), \qquad (2\leq i<n). \tag{9}$$

Thus, if the state transition matrix $T_0$ has already been computed, the last column of $T_1$ is known ($m=1$) except for $T_1(1,n)$. This element is obtained as follows. From (8)

$$T_{m-1}(1,1) = \frac{h^{m-1}}{(m-1)!} - a_1 T_m(1,n) \tag{10}$$

and

$$T_{m-1}(1,j) = T_m(1,j-1) - a_j T_m(1,n), \qquad (2\leq j\leq n). \tag{11}$$

In order to avoid the division operation by zero coefficients, the quantity $p$ is defined

$$p = \begin{cases} 0, & \text{if } a_1 \neq 0 \\ k, & \text{if } a_1 = a_2 = \cdots = a_k = 0, \quad a_{k+1}\neq 0 \\ n, & \text{if } a_1 = \cdots = a_n = 0. \end{cases} \tag{12}$$

If $p=n$ (integration of order $n$), (10) and (11) yield

$$T_1(1,n) = \frac{h^n}{n!} \tag{13}$$

And, if $p\neq n$, then from (10) and (11)

$$T_1(1,n) = \frac{1}{a_{p+1}}\left[\frac{h^p}{p!} - T_0(1,p+1)\right]. \tag{14}$$

Thus, if the state transition matrix $T_0$ has been computed, (9), (13), and (14) show that the forcing coefficients are determined.

To compute the state transition matrix itself, note that its elements satisfy the recursive relations [12]

$$\begin{aligned} T_0(i,1) &= -a_1 T_0(i-1,n) \\ T_0(i,j+1) &= T_0(i-1,j) - a_{j+1} T_0(i-1,n) \end{aligned} \quad \begin{cases} (2\leq i\leq n) \\ (1\leq j<n) \end{cases} \tag{15}$$

Therefore, only the first row of the transition matrix need be obtained from other considerations.

From the Cayley–Hamilton theorem, using the techniques of [7]–[10], we can prove that

$$T_0(1,j+1) = \frac{h^j}{j!} + h^j \sum_{k\neq n} \frac{\alpha_{j,k}}{k!}, \qquad (0\leq j<n) \tag{16}$$

where

$$\begin{aligned} &\alpha_{j,k} = \delta_{j,k}, \qquad (0\leq k<n) \\ &\alpha_{0,k} = -a_1 h^n \alpha_{n-1,k-1}, \\ &\alpha_{j,k} = \alpha_{j-1,k-1} - a_{j+1} h^{n-j} \alpha_{n-1,k-1} \end{aligned} \quad \Big\} \quad (k\geq n). \tag{17}$$

This gives the first row of the state transition matrix.

Combining (14) and (16)

$$T_1(1,n) = -\frac{1}{a_{p+1}}\left[h^p \sum_{k\geq n} \frac{\alpha_{j,k}}{k!}\right]. \tag{18}$$

Equations (9) and (18) express an important fact: *Once the state transition matrix is computed, only one extra division is required to obtain the forcing coefficients.*

## Procedure

The recommended procedure for the computation of all required coefficients for a single-input zero-order data hold is given below. This algorithm has been implemented as a Fortran subroutine called FACT[1] (acronym for Fundamental Algorithm for Computation of Transition) and verified on a variety of digital computers. Only the technique of terminating the series computation based upon the computer's least significant bit (utmost relative accuracy) appears to be correlated with computer hardware. For the trivial case of $n=1$, the state transition matrix is evaluated by exponentials.

1) Obtain $p$ from relationship (12).
2) For $j=0,\cdots,n-1$, do:
   a) if $j<p$, set $\phi_j = 0$;
   b) otherwise, calculate

$$\phi_j = h^j \sum_{k\geq n} \frac{\alpha_{j,k}}{k!}$$

until satisfied;

[1]Our Fortran version of FACT requires 500 memory cells and services any number of transfer functions.

c) Calculate $T_0(1,j+1)=(h^j/j!)+\phi_j$.

3) For $i=1,\cdots,n-1$, do:

a) calculate $T_0(i+1,1)=-a_1T_0(i,n)$;

b) For $j=1,\cdots,n-1$, do:

i) calculate $T_0(i+1,j+1)=T_0(i,j)-a_{j+1}T_0(i,n)$.

4) For the transient response only, terminate the algorithm. Only the state transition matrix is produced.

5) For $i=1,\cdots,n-1$, do:

a) Set $H(i+1)=T_0(i,n)$.

6) If $p=n$, calculate $H(1)=(h^n/n!)$ and terminate algorithm.

7) Otherwise, set $H(1)=-(\phi_p/a_{p+1})$.

As shown in (9) and (18), the $H$ vector produced by the algorithm is the last column of $T_1$ required to solve for the state.

## Computational Effort

Providing that tabular reference is made to inverse factorials, the stated algorithm requires $X$ multiplicative operations

$$X=n^2+2N(n-p)+3n+1 \tag{19}$$

where $n$ is the order of the system, $p$ is defined in relationship (12), and $N$ is the number of required terms for convergence of all the series (16).

## Series Properties

An estimate of the number $N$ of necessary terms for the convergence of the series as given by (16) depends on the $\alpha_{j,k}$ as given by relationships (17). It can be proven by a tedious induction that

$$\alpha_{j,k}=-\sum_{0\le i\le j} a_{j+1-i}h^{n+i-j}\alpha_{n-1,k-i-1}. \tag{20}$$

Furthermore, for an arbitrary constant $a\neq 0$

$$\alpha_{n-1,k}=[-\lambda]^{k-n+1}\Omega_k(\omega_1,\omega_2,\cdots,\omega_n) \tag{21}$$

where

$$\omega_i=\frac{a_{n+1-i}}{a^i} \tag{22}$$

and the coefficient size $\lambda$ is defined

$$\lambda\equiv ah. \tag{23}$$

The $\Omega_k$ are the generalized Lucas polynomials [13]. By defining the constant $a$ properly, various estimates of $N$ can be derived from bounds on these polynomials. Particularly convenient bounds on the $\Omega_k$ can be derived if the $\omega_j$ are bounded by

$$0<|\omega_j|\le 1, \qquad (1\le j\le n). \tag{24}$$

This inequality will be satisfied for all coefficients if $a$ is defined[2]

$$a=\underset{1\le j\le n}{\text{Maximum}}\left[|a_j|^{1/n+1-j}\right]. \tag{25}$$

By (21)–(23), $N$ will depend on $n$, the parameter $\lambda$ (which measures the size of the coefficients), and the $\omega_j$ (which measure the relative distribution of coefficients for fixed $\lambda$). It will be shown statistically that for most practical systems, with $a$ given by (25), the effect of the distribution of coefficients on convergence is secondary.

In order to determine the influence of $\lambda$ upon convergence, Monte Carlo methods were used. A range of $0<\lambda\le\lambda_{max}$ was selected for the statistical sample. The range of $\lambda$ is related to the sum of the system pole radii defined by

$$p_{sum}=\sum_{1\le i\le n}|p_i| \tag{26}$$

where the $p_i$ are the system poles. It can be proven that for any system the coefficient size $\lambda$ satisfies

$$\lambda\le\lambda_{max}\equiv hp_{sum}. \tag{27}$$

For each $\lambda_{max}>0$ the system

$$a_k=\binom{n}{k-1}\left[\frac{\lambda_{max}}{nh}\right]^{n+1-k} \tag{28}$$

which has all poles real and equal to $\lambda_{max}/(nh)$ satisfies

$$\lambda=\lambda_{max}. \tag{29}$$

Thus any class of systems with maximum pole radii sum $p_{sum}$ has a maximum $\lambda=\lambda_{max}$ as given by (27). The converse is not true: if $0<\lambda<\lambda_{max}$, then it does not follow that the pole radii sum is bounded by the corresponding $p_{sum}$.

The dependence of $\lambda_{max}$ on pole size can be given the following geometric interpretation. Define the *average pole radius*

$$p_{av}\equiv p_{sum}/n \tag{30}$$

then by (27)

$$\lambda_{max}=nhp_{av}. \tag{31}$$

Thus for systems of average pole radius bounded by some fixed value and for fixed $h$, the larger the system order the larger the possible range of $\lambda$. The range $0<\lambda\le\lambda_{max}=10$ was selected as a reasonable range of $\lambda$ in terms of pole size. For example, for $n=5$, if the average pole radius is bounded by 20 ($p_{sum}=100$), then $0<\lambda\le 10$ for the system and it is included within the range considered in the analysis below. Note that the largest pole may be much greater than 20; only the average of the pole radii need be bounded by 20. Furthermore, the chosen range may even include some systems for which the average pole size is greater than 20. Although no general statement can be made for these latter systems, there exist those with arbitrarily large poles for which $\lambda$ lies within the stated range.

A total of 20 000 sets of random coefficients was obtained for each $n$ divided among $\lambda$ intervals of 0.5. The fifth-order system, which is presented here in detail, was determined to be representative. For a given $\lambda$, the required number of terms $N$ is a statistical variable. In Fig. 1 the average (dotted line) and plus or minus one standard deviation

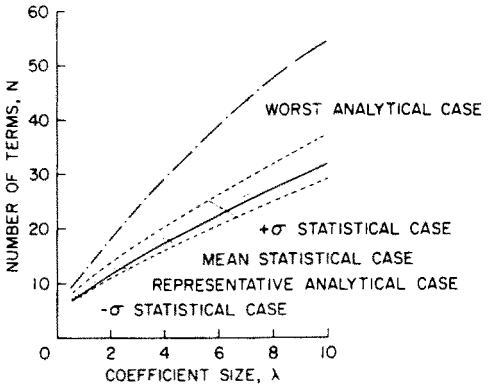


Fig. 1. Distribution for fifth-order systems.

[2] $N$ is zero if $a$ vanishes.

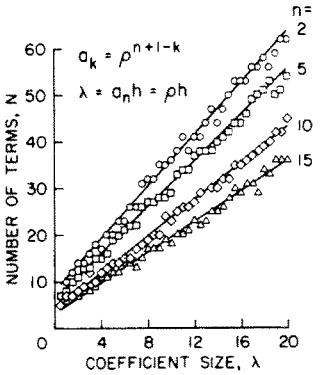


Fig. 2. Influence of system order on convergence.

(dashed lines) are presented. $N$ is seen to statistically vary but a small percentage from its mean value. Thus $\lambda$ is the primary parameter for determining convergence. In Fig. 1 there appears an alternately dotted and dashed line; this will be seen to represent the worst possible distribution of coefficients. Also shown in Fig. 1, the solid line will be seen to represent typical systems of a wide class.

Specialization to a class of systems permits various analytical estimates of $N$. One such class contains those coefficients which satisfy the inequality

$$|a_k| \leq |a_{k+1}||a_n|, \qquad (1 \leq k < n). \tag{32}$$

For example, it can be proven in general that if all poles lie in a 60° sector about the negative (or positive) real axis, this inequality holds for the coefficients. Thus inequality (32) defines a large class of coefficients. For this class $a = |a_n|$, permitting a simple evaluation of $a$.

An absolute worst convergence case for all coefficients is contained within this class. It can be deduced from (21) as follows. Select $\rho > 0$ arbitrarily and define

$$a_j = (-1)^{n-j}\rho^{n+1-j}. \tag{33}$$

Then $a = \rho$. It can then be proven that for each $k$, $\Omega_k > 0$ and assumes a maximum value at $\omega_j$ corresponding to $a_j$ of (33) for all $\omega_j$ satisfying (24), which includes all coefficients $a_j$. In view of (16), (20), and (21), it can be expected that the distribution (33) will give the slowest convergence of the series.

A survey of all the statistical data used to generate Fig. 1 shows that the system (33) indeed bounds the data.

Within the class of systems satisfying inequality (32) a nominal convergence case can be extracted. Such a representative system is defined

$$a_k = a_n^{n+1-k}, \qquad (1 < k \leq n). \tag{34}$$

The number of terms for convergence of this system versus $\lambda$ is given in Fig. 1 as a solid line. It lies between the best and worst statistical cases obtained; similar results were obtained for other values of $n$. The conclusion to be drawn from Fig. 1 is that the convergence properties of the system given by (34) are representative of most systems of practical interest.

Using the system (34) as a practical convergence case, plots of $N$ versus $\lambda$ are presented in Fig. 2. It is seen that the larger the order of the system the faster the convergence. Since the system (34) is considered typical of most systems, Fig. 2 gives a convenient means to evaluate the number of terms $N$ for convergence used in (19) in order to determine the efficiency of the FACT algorithm. For a class of systems, the faster convergence for larger $n$ at fixed $\lambda$ can be interpreted geometrically in terms of the poles. For stable systems satisfying inequality (32) we have

$$\lambda = ah = h \sum_{1 \leq i \leq n} |\mathrm{Re}(p_i)|. \tag{35}$$

Thus smaller poles tend to accelerate convergence.

By formulating the convergence criterion in terms of $\lambda$, we have identified the dependence of convergence on the transition interval used and on the size of the system poles. The latter follows since the parameter $a$ used in the definition of $\lambda$ in (23) serves as a measure of the size of the poles of the system. We have seen from (31) that a given $\lambda$ and hence a given $a$ places a lower bound on the average pole radius. For larger $a$ we must have larger poles. Furthermore, for the system (34) it can be verified that the poles coincide with $a$ times the complex roots of unity excluding the root at $s = 1$. Thus for this important system $a$ is the pole radius to the origin. Finally, for more general systems satisfying (32), by (34) $a$ measures the average real part of the poles. For stable systems as the poles move away from the origin, $a$ increases and the convergence is slower (for fixed $h$). However, in practice, a limit is placed on the size of $\lambda$ in order to observe the system frequency content with the transition interval $h$ used. For large values of $a$, which implies large poles, the parameter $h$ must be reduced. Thus the definition of a practical value for $\lambda$ produces a reasonable value for the number of terms required for convergence of the series.

## Accuracy

The accuracy of the FACT algorithm has been established by comparisons with elements obtained from double-precision matrix series summations. Evaluations were made with statistical samples of 2000 points for each system order $n = 2$, 5, and 10 for $0 \leq \lambda \leq 10$ divided among intervals of 0.5. The measure of absolute error utilized was the norm of the difference between transition matrices and forcing function vectors obtained by these two techniques.

The average relative error was typically less than one-hundredth of one percent, with a maximum error of one tenth of one percent.

## Conclusions

We find that the position of the coefficients plays the dominant role in the series convergence properties. Thus a large $a_n$ will slow convergence far more than a large $a_1$. The critical parameter is $\lambda$; the distribution of the coefficients has a secondary effect on convergence. Also, the larger the order of the system, the more rapid the convergence. This aids somewhat in counteracting the necessary $n^2$ multiplications.

The FACT algorithm produces a dramatic reduction in the computational requirements, providing only that the coefficient matrix be in standard companion form. Only one extra division is required over the computational effort needed for the state transition matrix in order to compute the forcing-function coefficients.

This analysis assumed a zero-order hold on the input. The results obtained here can be easily extended for the higher order hold cases.

## References

[1] E. J. Davison, "The numerical solution of $\dot{X} = A_1X + XA_2 + D$, $X(0) = C$," *IEEE Trans. Automat. Contr.* (Corresp.), vol. AC-20, pp. 566-567, Aug. 1975.
[2] R. E. Kalman, "Mathematical description of linear dynamical systems," *SIAM J. Contr.*, ser. A, vol. 1, pp. 152-192, 1963.
[3] W. G. Tuel, Jr., "On the transformation to (phase variable) canonical form," *IEEE Trans. Automat. Contr.* (Corresp.), vol. AC-11, p. 607, July 1966.
[4] J. S. Frame, "Matrix functions and applications," pt. V, *IEEE Spectrum*, p. 104, July 1964.
[5] F. Neuman and J. D. Foster, "Investigation of a digital automatic aircraft landing system in turbulence," NASA TN D-6066, pp. 10-17, Oct. 1970.
[6] R. E. McFarland, "A standard kinematic model for flight simulation at NASA-Ames," NASA CR 2497, p. 28, Jan. 1975.
[7] W. E. Thomson, "Evaluation of transient response," *Proc. IEEE* (Lett.), vol. 54, p. 1584, Nov. 1966.
[8] I. Kaufman, "Comment on 'Evaluation of transient response,'" *Proc. IEEE* (Lett.), vol. 58, p. 143, Jan. 1970.
[9] S. Ganapathy and A. Subba Rao, "Transient response evaluation from the state transition matrix," *Proc. IEEE* (Lett.), vol. 57, pp. 347-349, Mar. 1969.
[10] V. Purna Chandra Rao, "Comments on 'Transient response evaluation from the state transition matrix,'" *Proc. IEEE* (Lett.), vol. 58, p. 814, May 1970.
[11] C. L. Krouse and E. D. Ward, "Improved linear system simulation by matrix exponentiation with generalized order hold," *Simulation J.*, vol. 17, pp. 141-146, Oct. 1971.
[12] I. Kaufman, "Evaluation of an analytical function of a companion matrix with distinct eigenvalues," *Proc. IEEE* (Lett.), vol. 57, pp. 1180-1181, June 1969.
[13] R. Barakat and E. Baumann, "$M$th power of an $N \times N$ matrix and its connection with generalized Lucas polynomials," *J. Math. Phys.*, vol. 10, pp. 1474-1476, Aug. 1969.